**BIOLOGICAL CRITERIA**
Technical Guidance for Survey Design and Statistical Evaluation of Biosurvey Data

# APPENDIX A Basic Statistics and Statistical Concepts

Certain specific features of a data set are characterized by descriptive statistics. Of these measures, the center, or central tendency of a set of data, is probably the most important. Among the candidate statistics for central tendency are the mean, median, mode, and geometric mean. Once the center of a data set is described, the next important feature is the data distribution: the spread, dispersion, or scale. Among the candidate estimators of dispersion are range, standard deviation, and interquartile range. These two characteristics of a data set, central tendency and dispersion, are the most common descriptive statistics. Other characteristics, such as skewness and kurtosis, are occasionally important. The examples that follow illustrate the choice of descriptive statistics.

## Measures of Central Tendency

Probably the single most useful way to summarize a data set is to indicate the center of the sample. "Center" suggests the vague notion of the middle of a cluster of data points or perhaps the region of greatest concentration. Since samples of data exhibit a variety of distributions when plotted as bar graphs (histograms), it is not possible to define the center unambiguously. As a result several statistical estimators can serve as candidates for determining central tendency or location, and each candidate has advantages and disadvantages for the task at hand.

### *Mean*

The arithmetic mean, or simply, the mean — the sum of all data values divided by their number — is the most frequently used central tendency estimator. It is so commonly used that scientists often lose sight of the true reason for calculating descriptive statistics. In some cases, the mean is calculated as the central tendency, though another central tendency statistic would be better.

The arithmetic mean ($x$) is the sum of the observations ($x_i$) divided by the number of observations ($n$):

$$x = \frac{\sum x_i}{n} \qquad \textbf{(A.1)}$$

Each observation contributes its magnitude to the sum of the observations and hence to the mean. For symmetric distributions (like the normal bell-shaped or Gaussian distribution), the mean calculated from a sample of data (the sample mean) often comes quite close to the center, or peak, of the histogram for that sample. However, biological data are often not symmetrically distributed. The extremely high or extremely low observations characteristic of skewed (nonsymmetrical) data distributions pull the mean in the direction of the skew; a few extremely high observations can pull the mean away from the bulk of the observations and toward the few high data points. In those situations, a more resistant estimator, such as the median or the mode, may be preferred.

### *Median*

The median is the value of the middle observation when data are arranged in order of size — from lowest to highest value. The median is therefore known as an "order statistic" since it is based on an ordering or ranking of observations. When the total number of observations is an even number, leading to two middle values, the median is then the average of the two middle values.

The "order" of the median observation is

$$\text{Median Observation} = (n + 1)/2 \qquad \textbf{(A.2)}$$

The effect on the median of all but the middle-ranking observations is simply to hold a place in the ranking so that outlying observations do not pull the median toward the extremes. The median is resistant to the influence of any particular observations; therefore, it is a good statistic to use when the histogram is skewed or unusually shaped.

### *Trimmed Mean*

The trimmed mean is the mean value from a subsample of the original sample. The subsample is formed by symmetrically trimming a small percentage of the data points from either end of the ordered observations. For example, a 10-percent trimmed mean is calculated from the subsample remaining after the highest and lowest 10 percent of the observations are removed from the set. At the extreme, the median is the trimmed mean with all but the middle observation removed.

The trimmed mean is an efficient indicator of central tendency if censoring has occurred or if a few outlying observations are found in the data. Here, censoring refers to data points reported as "below detection limits." If 15 percent of the data points are be-

low detection limits, then a 15-percent trimmed mean estimator (involving 15 percent trimming from each end) should result in less bias than the arithmetic mean, the estimator based on all uncensored observations.

### *Mode*

The mode is the value in the sample that is most frequently observed; it can be used for discrete or categorical data. If no value is repeated more than once, as is possible for biological data on a continuous scale, the mode will not be a useful estimator of central tendency. Alternatively, if a histogram is used to represent the data, then the mode is defined as the range of values associated with the tallest bar on the histogram. The mode is a good estimator for central tendency because the most frequently observed value is usually near the center of the distribution. The histogram will indicate visually whether the mode actually does correspond with the center of the sample.

### *Geometric Mean*

The geometric mean is a reasonable measure of central tendency for a set of data that exhibit a lognormal distribution. It is the antilog of the mean of logarithmically transformed data. The lognormal data distribution is skewed in the original units of measurement, but normal (Gaussian) when the original measurements are log-transformed. Several investigators suggest that the lognormal distribution is a good probability model for concentration data on environmental contaminants. Data sets described by the lognormal distribution have a few high values that are somewhat extreme from the bulk of the observations.

The geometric mean may be calculated in two ways:

$$Geometric\ Mean = anti\log\left(\frac{\sum \log(x_i)}{n}\right) \qquad \textbf{(A.3)}$$

or:

$$Geometric\ Mean = \left[\prod x_i\right]^{\frac{1}{n}} \qquad \textbf{(A.4)}$$

where $\prod x_i = x_1 \cdot x_2 \cdot x_3 \cdot \ldots \cdot x_n$.

## Measures of Dispersion

If central tendency measures are not used to summarize a data set, then measures of dispersion or spread will be used instead. Dispersion in a data set refers to the variability in the observations around the center of the distribution. Good measures of dispersion will be obtained from symmetric distributions. Asymmetry, or skew, will affect the estimate of dispersion so that it overestimates spread in the shorter tail of the data distribution (while underestimating it in the longer tail). A transformation (e.g., logarithm) should be considered in cases of asymmetry in order to create a symmetric distribution. Statistics are then calculated on the basis of the transformed metric.

### *Standard Deviation*

The most commonly used statistic for dispersion is the standard deviation. In fact, the standard deviation, like the mean, is used so often that it is sometimes thought to be the equivalent of dispersion. It is, however, a measure of variability that represents the average distance of the data from the mean; and, like the mean, it is strongly affected by extreme values. Thus, the standard deviation for a distribution of data with a long tail to the right is inflated by the values at the extreme right. Investigators may prefer to create a symmetric distribution before calculating the standard deviation.

For a sample, the sample variance ($s^2$) is

$$s^2 = \sqrt{\frac{\sum (x_i - x)^2}{n-1}} \qquad \textbf{(A.5)}$$

and the sample standard deviation ($s$) is the square root of the variance ($\sqrt{s^2}$).

### *Absolute Deviation*

The standard deviation is based on squared error; squaring the deviation between a data point and the sample mean increases the influence of the largest and smallest observations on the estimate of deviation. The absolute deviation can be calculated to reduce the influence of outliers on the dispersion statistic. To arrive at the absolute deviation, the mean (or median) is first estimated, and then the absolute value of the difference between the mean or median and each data point is calculated. The mean or median of these absolute deviations is then calculated as the mean or median absolute deviation.

### *Interquartile Range*

Since the standard deviation is unduly influenced by extreme observations in both symmetric and asymmetric distributions of data, a resistant alternative to the standard deviation (as the median is to the mean) is needed for situations in which the data are skewed but transformation is undesirable. Fortunately a good alternative exists — the interquartile range: the range that includes the central 50 percent of all observations in the set. The interquartile range, like the median, is based on order statistics; thus, it is unaffected by the magnitude of the extreme observations in either tail. It is calculated as the difference between the

observation at the 75th percentile (upper quartile) and the observation at the 25th percentile (lower quartile):

Lower quartile rank order =
(1/2)(1 + median rank order)

Upper quartile rank order =
(1/2)(1 + n + lower quartile rank)

Interquartile range (I) =
lower quartile value – upper quartile value.

### *Range*

Range is an easily determined and therefore frequently cited measure of dispersion. The range is simply the maximum value minus the minimum value. Since it is clearly affected by the magnitude of the observations at either extreme, the range should not be relied on as the sole indicator of variability. Nevertheless, it is often informative to list the range along with another dispersion statistic.

## Resistance and Robustness

In a number of scientific fields, particularly those that depend on observational (as opposed to experimental) data, errors of measurement and natural variability are apt to result in empirical distributions (histograms) with occasional outliers and shapes that are more spread-out than the normal density function. This result, which is fairly common in water quality studies, makes robustness and resistance important considerations when choosing statistics to summarize data. In some situations, of course, the outliers, rather than central tendency and dispersion, will be the focus of the study.

A resistant estimator is one that is insensitive to data points that are quite different from the rest of the data (i.e., outliers). A robust estimator is one that performs well (efficiently), even if an assumption concerning the underlying probability model is wrong. For central tendency, the mean is neither resistant nor robust. The median is resistant to outliers but not robust since it is not as efficient as other options (i.e., it is subject to large standard error). The trimmed mean and so-called M-estimators (Hampel et al. 1986) are both resistant and robust.

The most commonly used measure of dispersion, the standard deviation (or variance), is nonresistant (highly sensitive to outliers) and not robust because squaring the deviation emphasizes deviant data points. The absolute deviation and the interquartile range are more resistant but not highly robust.

Resistance and robustness provide a measure of insurance against features of the sample data that may yield a summary estimate that is not representative of the data set as a whole. A robust and resistant estimator is not the best choice if, for example, there are no outliers and the sample is an exact normal density function. However, if outliers do occur, and samples are not normal (or lognormal), then robust and resistant estimators of center and dispersion are wise and safe choices that will help investigators avoid faulty inferences.

## Graphic Analyses

It is good practice in statistical analysis to begin with various displays of the raw data. That is, before descriptive statistics are calculated from a data set, and before analyses such as hypothesis testing and linear (regression) model building occur, it is wise to look at empirical graphs. The graphs recommended for this task help the investigator identify the need to transform the data before conducting the statistical analysis.

Most procedures in statistics (e.g., regression analysis, hypothesis testing) derive summary values (e.g., mean, trimmed mean) from a data set. If the inferences drawn from statistical procedures are to be valid for the entire data set, then the summary statistics must represent the entire set. Graphic displays guide the choice of any necessary manipulations of the data set and help assure the selection of appropriate summary statistics. The examples presented here underscore the importance of displaying the data at the beginning of a statistical study.

Graphs can also be useful during the course of a statistical study. For example, bivariate scatter plots help scientists select independent variables for a regression equation, and scientists will often wisely choose to present the results of a statistical analysis in graphic form. Conclusions are often most effectively conveyed through graphs.

### *Histograms*

Perhaps the most fundamental level of study is an analysis of data on a single characteristic. Assume, for example, that an aquatic biologist has a data set for species richness from a stream study and now desires to summarize this information. The biologist could calculate the trimmed mean and median absolute deviation of the sample; alternatively, she could calculate other statistics representing central tendency and dispersion. To determine which of these statistics are most useful, the biologist should first look at a plot of the data. The histogram is often used to display data representing a single characteristic (such as IBI).

For example, suppose that the index of biotic integrity in Table A.1 has just been determined for a particular stream from headwaters to mouth, and the

**Table A.1.—IBI data for a particular stream.**

| | IBI | DATA | |
|---|---|---|---|
| 12 | 25 | 33 | 56 |
| 12 | 24 | 34 | 58 |
| 14 | 26 | 35 | |
| 15 | 24 | 36 | |
| 16 | 24 | 35 | |
| 22 | 27 | 38 | |
| 24 | 23 | 41 | |
| 23 | 28 | 42 | |

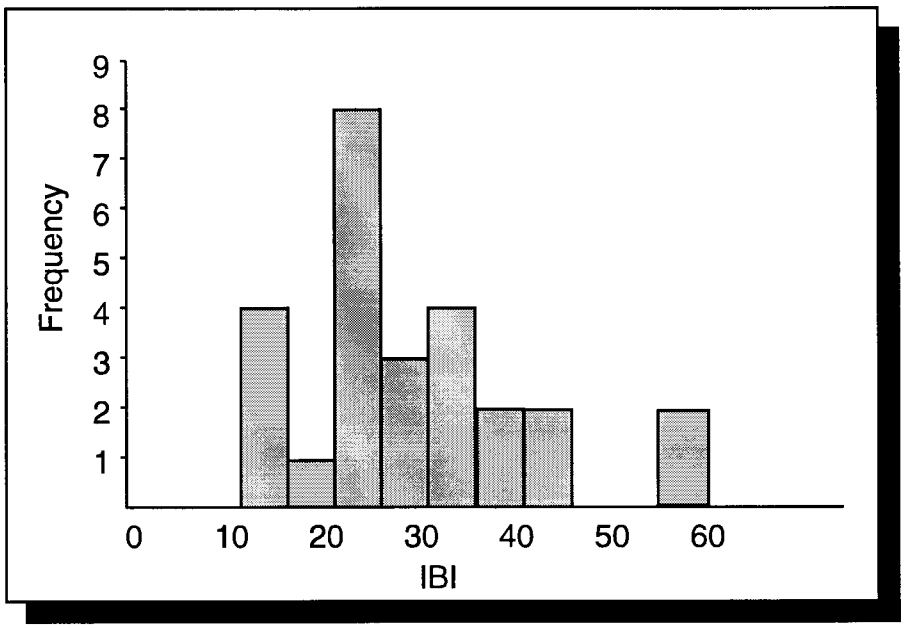


Figure A.1.—Histogram of IBI data for a particular stream.

biologist wants to picture the biotic integrity of this stream. As a first cut, the histogram in Figure A.1 is plotted. To construct the histogram, the biologist must first divide the range into equal intervals. In Figure A.1, the range is approximated by 10 to 60 (actually it is 12 to 58) and is divided into intervals of 5 units. For each interval, 11 to 15, 16 to 20, and so on, the height of the bar represents the number of data points that lie within that interval. So there are four IBI data points that lie within 31 to 35 and eight within 21 to 25. Thus, the bar for the 21 to 25 interval is twice the height of the 31 to 35 bar.

What does the histogram tell us about this stream? Basically, it provides us with a visual image of the distribution of the sample. In specific terms, it means that we can quickly see such things as the location of the center of the sample, amount of dispersion, extent of symmetry, and the existence of outliers in the sample. Outliers need not be errors or aberrations; they are simply "set apart" from the bulk of the observations. The reasons why they are set apart may be of particular interest in some studies.

In Figure A.1, the center may be visually associated with the highest bar (mode) at 21 to 25, or it may be identified as a middle value (median) around 30. Dispersion could perhaps be characterized by stating that the range is 12 to 58, and almost 60 percent (actually 15 to 26) of the data points lie between 20 and 35. The histogram is not symmetric, however, and one might want to check on the validity of the two outlying observations on the extreme right.

The picture created by the histogram is of considerable value in the selection of descriptive statistics. Some care should be observed in the construction of the histogram, however. With changes in interval size, the histogram can assume different shapes that may affect the inferences. For example, the IBI data in Figure A.2 are plotted using an interval size of 10 units. On that scale, the two highest data points no longer appear as outliers. In contrast, the two-unit intervals in Figure A.3 give the impression of possible outliers on both the right and left extremes of center. It is probably good practice to scale the histogram so that the observations are neither too aggregated (as in Figure A.2) nor too spread out to permit reasonable inferences to be drawn.

Thus, the histogram provides an impression of the extent of symmetry in the sample. Symmetry in a data set is a desirable attribute for two reasons. First, it often means that one can characterize the sample as having a distribution with a shape similar to one of the symmetric distributions (e.g., the normal distribution), which is often assumed to be an underlying model in statistical inference. Stating, for example, that a sample approximates the normal distribution conveys useful information. Beyond that, symmetry implies that common descriptive statistics are clear: central tendency refers to the center of symmetry, and dispersion characterizes variability without skew.

Therefore, it may be useful to apply a transformation, if necessary, to create symmetry in an asymmetric data set. Continuous concentration and

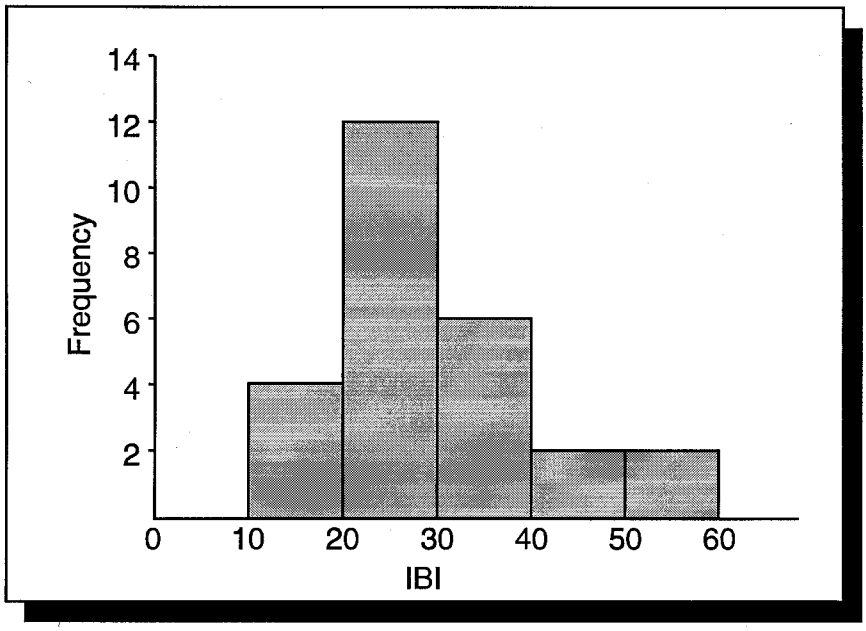


**Figure A.2.—Histogram of IBI data with 10-unit intervals.**

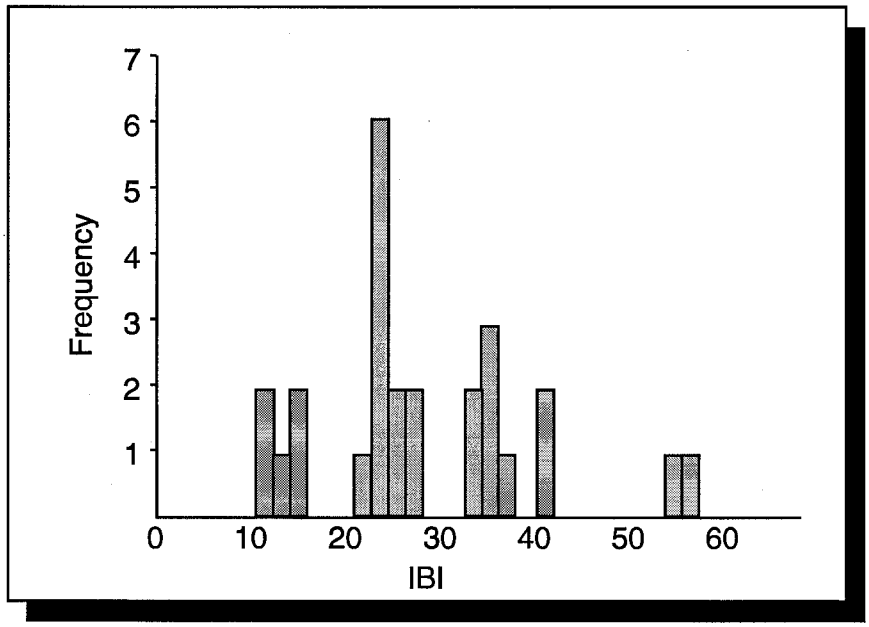


**Figure A.3.—Histogram of IBI data with two-unit intervals.**

density data often yield a skewed-right histogram, since all values are positive and a few relatively high values frequently occur. This description characterizes the lognormal distribution, which is often used to describe biological data sets. To check for lognormality, the logarithmic transformation is applied to data, and a histogram of the transformed data is plotted. Comparison of this histogram with a normal density function provides a rough indication of lognormality; formal tests also exist (e.g., the Kolmogorov-Smirnov test). These tests can be found in many introductory statistics books.

To illustrate how a transformation can change the shape of a histogram, the IBI data from Table A.1 were log-transformed, and histograms of the log IBI are presented in Figures A.4 and A.5.

Compare Figure A.2 with Figure A.4; in the first figure, the distribution of IBI appears skewed-right, whereas in the second figure, the distribution of log IBI is skewed left. Figure A.5 appears approximately symmetric and normal, which suggests that the logarithmic transformation may be a good idea if these attributes are desired. Note that the data points at the extreme right in the original IBI metric no longer appear to be outliers in the log-metric. This result is the effect of the logarithmic transformation — it has spread out the low values and squeezed in high values. Having studied the histograms of this data set, we can now determine which statistics are most appropriate for our data summary.

## *Stem and Leaf Displays*

An alternative and often informative version of the histogram is the stem and leaf display. Developed by Tukey (1977), the stem and leaf plot provides the shape of a histogram and the data's numeric values simultaneously. For example, the stem and leaf display for the IBI data set in Table A.1 is presented in Figure A.6. Note that its shape is nearly the same as the histogram in Figure A.1.

To construct the stem and leaf diagram, first choose the stems digit and the interval width. In Figure A.6, the stem is assigned to the tens digit, and the interval width is one-half of 10, or five. The values for the stem are placed to the left of a vertical line. Each stem digit is repeated in Figure A.6 because the interval width is five units. Thus the first tens stem covers 0-4 and the next covers 5-9. On the right side of this

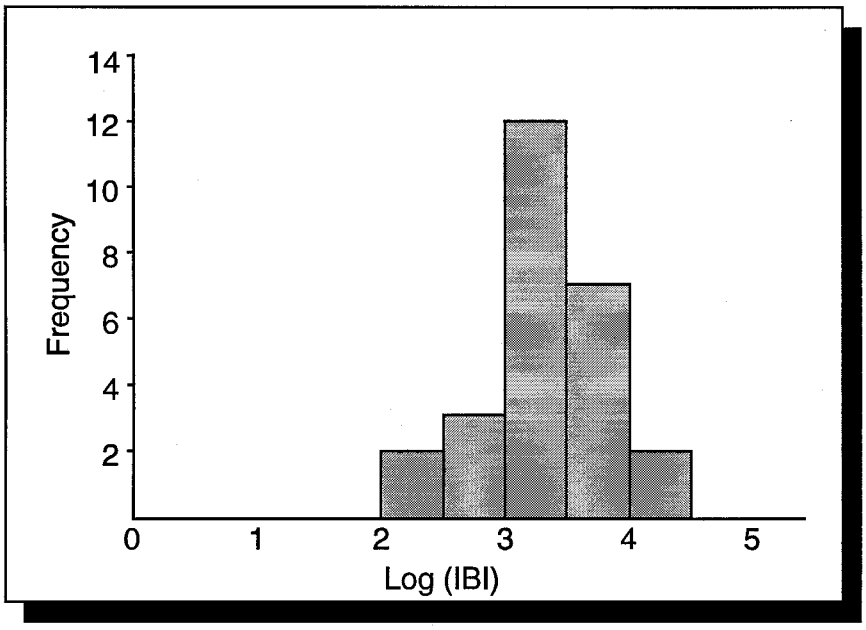


**Figure A.4.—Histogram for log(IBI).**

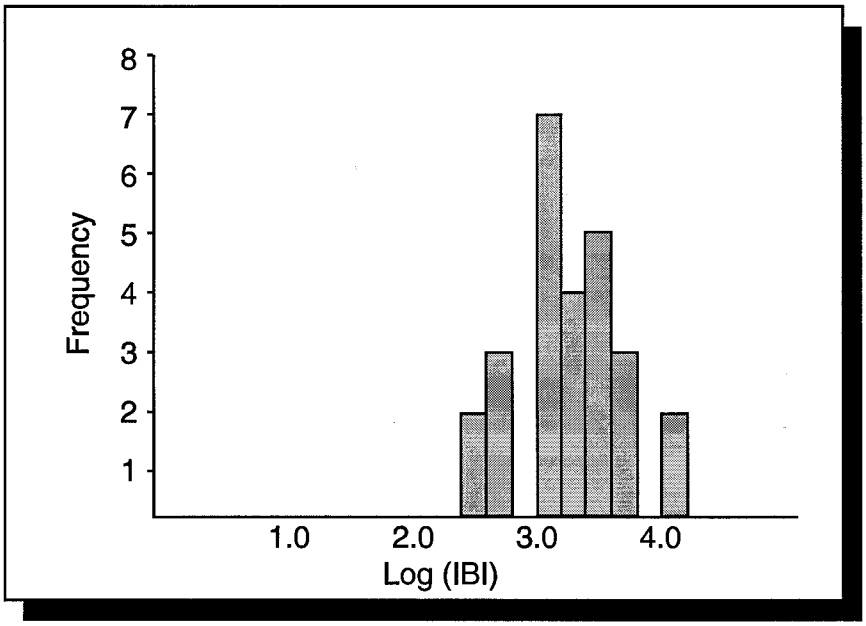


**Figure A.5.—Histogram for log(IBI): Alternative scale.**

line, the "leaves" are written. For each data point, the leaf is the next digit lower in value than the stems digit. Since the stems in Figure A.6 are composed of the tens digit, the leaves are made up of the units digits. Each observation contributes one leaf to the row containing its stem. For the IBI data points in Table A.1, the first observation (12) results in a 2 (the units digit) placed in the row for the first tens stem (covering IBI from 10 to 14). The IBI of 16 results in a 6 (the units digit) placed in the row for the second tens stem (covering IBI from 15 to 19), and so on.

The primary advantage of the stem and leaf display (over the histogram) is that it contains information on the numeric values in the data set (and still provides information on the shape of the sample distribution). There may be advantages to this combination, particularly when the data are displayed for presentation purposes. Tukey (1977) describes several variations of the stem and leaf display, including an interesting way to look at covariation in bivariate data.

## *Box and Whisker Plots*

Investigators often need to compare two or more samples of the same characteristic (e.g., samples of the IBI for the same waterbody for two or more years). This comparison may be purely statistical, perhaps using hypothesis testing. Alternatively, a graphic method could be used —

| | |
|---|---|
| 1 | 224 |
| 1 | 56 |
| 2 | 2334444 |
| 2 | 5678 |
| 3 | 34 |
| 3 | 5568 |
| 4 | 12 |
| 4 | |
| 5 | |
| 5 | 68 |

Figure A.6—Stem and leaf display.

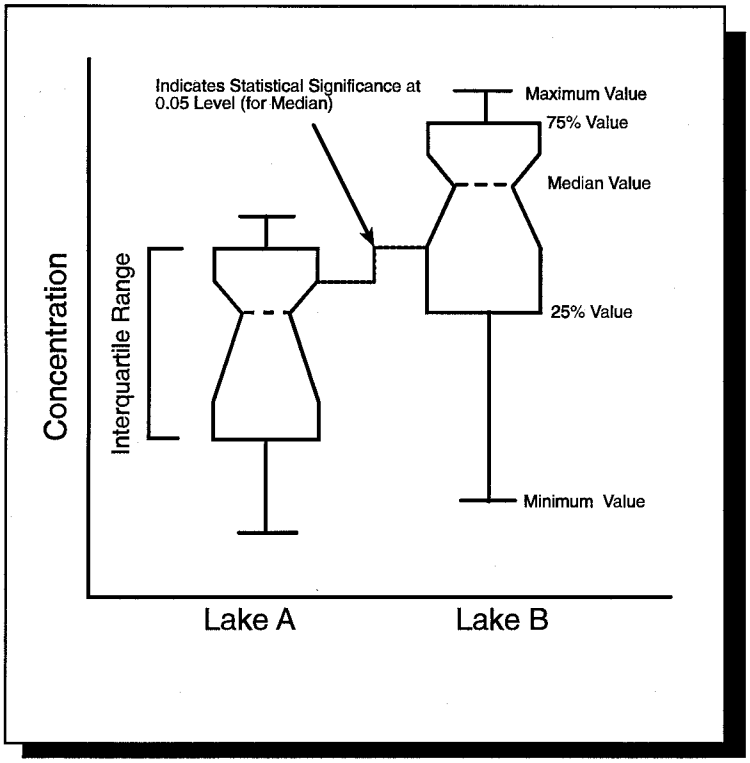


**Figure A.7.—Box and whisker plots.**

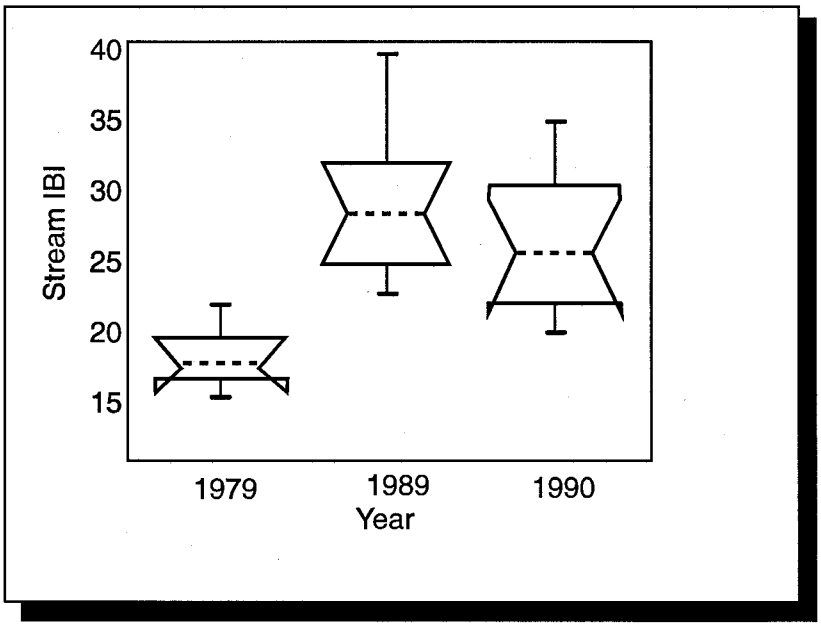


Figure A.8.—Stream IBI box plots.

Figure A.7 shows the basic structure of the box plot. For clarification, note that the "statistical significance of the median" on Figure A.7 refers to the degree of vertical overlap of the notch or indention in one box with the notch in another box. If the notches do not overlap vertically, then the medians may be considered significantly different at approximately the 0.05 level.

Box plots are based on order statistics which, like the median, are calculated by ranking the observations from lowest to highest. Box plots can be used to convey information on the sample median; dispersion, as conveyed by the range and the interquartile range; skew, as conveyed by the symmetry in the shape above and below the median; relative size of the data set, as conveyed by the width of the box; and statistical significance of the median.

Figure A.8 shows three sample box plots for stream IBI data for 1979, 1989, and 1990. The box and whisker plots in Figure A.8 provide a substantial amount of information on IBI during the years of sampling. First, it is apparent that IBI has increased since 1979, as there is little vertical overlap of the 1979 box plot with the other two. This conclusion is further supported by the lack of vertical overlap in the 1979 notch with the other two notches. In contrast, while the medians for 1989 and 1990 differ, they are not significantly different (0.05 level) and the samples (boxes) overlap considerably. None of the years exhibit substantial skew in the sample data. The 1989 data are skewed the most, based on the relative symmetry of the box and whiskers around the median.

Box plots are helpful as diagnostic tools and as a method of demonstrating conclusions about samples following the completion of a statistical study. Tukey (1977) and Reckhow (1979) describe several interesting applications.

erhaps one that provides both pictorial and statistical comparison. One such model is the box and whisker plot, which is available in many statistical software packages for the microcomputer.

## *Bivariate Scatter Plots*

Many statistics (e.g., correlation coefficients) and many statistical methods (e.g., regression analysis) are fundamentally concerned with relationships between pairs of variables. Without doubt, the best

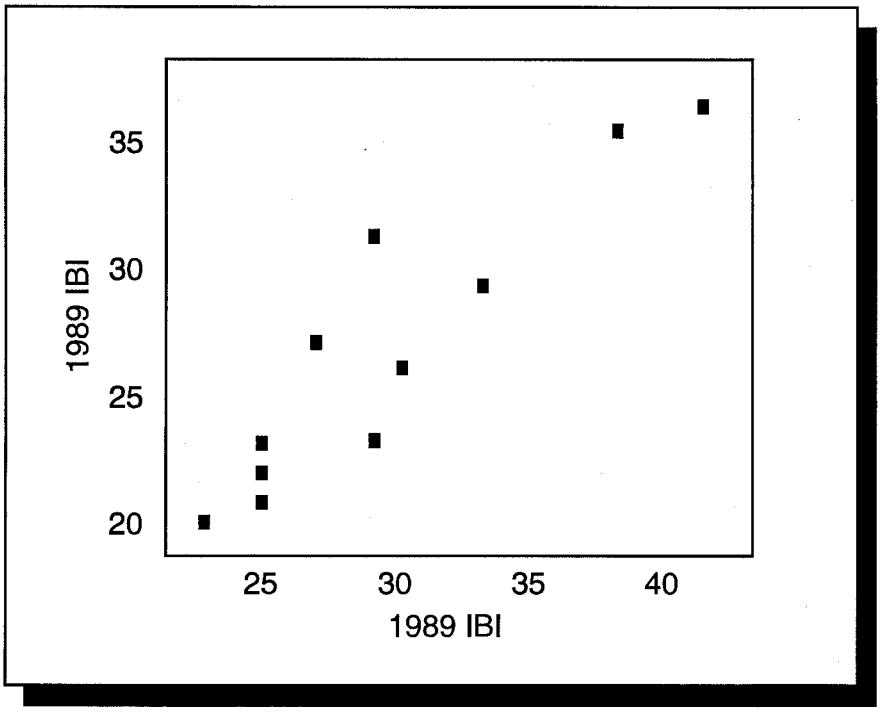


**Figure A.9.—IBI bivariate plot for 1989 and 1990 data.**

A second topic of interest for bivariate samples is the presence or absence of outliers. Outliers have no universally accepted objective definition; rather, the term is used here to identify observations that stand apart from a cluster of points. We are concerned about outliers because they are apt to have excessive influence on nonresistant statistics like the mean, variance, sample correlation coefficient, and OLS regression coefficients. Bivariate plots are valuable for outlier identification and may suggest approaches (e.g., transformation) for correction. In Figure A.9, the two highest values probably would not be considered outliers, since they are compatible with the pattern exhibited in the rest of the data and not substantially separated from those data.

way to examine a relationship between pairs of variables (a bivariate relationship) is through a scatter plot.

In Figure A.9, a bivariate scatter plot is presented for the 1989 and 1990 IBI data for a particular stream. From the plot, we can examine the distribution of data for each variable separately and for the two variables together. For example, we can see from Figure A.9 that two relatively high observations tend to stand apart from the rest of the data, particularly in the horizontal direction. As might be expected, there is an approximately linear correlation between the IBI estimates for successive years.

Two characteristics of a bivariate sample are often of interest in statistical studies. First, the biologist may be interested in the pattern or shape (e.g., linearity or nonlinearity) of a relationship. Linear relationships are often desirable for ease of analysis; correlation analysis and ordinary least squares (OLS) regression provide measures of the strength of a *linear* relationship. If the bivariate relationship is nonlinear, it is possible that a transformation can be applied to make it linear, or a nonlinear model may be used. Without question, the scatter plot is the most important diagnostic device for evaluating linearity, and it is often quite helpful in selecting a transformation.